Machine Learning Methods for Web Documents

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Author: Ing. Josef Katrňák

Supervisor: doc. Ing. Radek Burget, Ph.D.

AIMS OF THE THESIS

• This work aims to use **machine learning techniques** for the **classification** of specific parts of **web page** content.

• **Visual representation** of web pages serves as input for **training** of machine learning models.

• The model architecture is based on **graph neural networks.**

• The advantage of the proposed and implemented approach is **information extraction** independent of the structure and language of a web page.

ABSTRACT

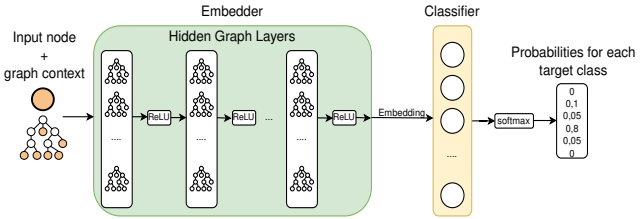
• The page is first converted to a **graph model** where each node contains a number of **visual features**.

• **Leaf nodes** can contain a required information and are the target of classification.

• Target nodes with the context of a graph enter the **Embedder**, which consists of graph neural layers.

• From the Embedder comes an **embedding** that represents a node relative to the **desired task** and **graph context**.

• The embedding is classified by the **Classifier**, resulting in **probabilities** that indicate the degree to which the node **contained a information** of interest.



MOTIVATION

• Identification and storing of **important data** from the website.

• Use of **visual and spatial** information.

• Quick **retrieval of key information** from any web page in a given category.

• **Save time** when searching for information.

• Possibility of **further processing** of automatically obtained information.

RESULTS

• The best trained models achieve an accuracy of **98.38%** with an F1 score of **0.9837**.

• The success rate of finding the desired information is up to **97.83%**.

• Model was compared with other work using a predictive accuracy metric:

	Text Features	Name	Price	Image	Add To Cart	Go To Cart	Average
FreeDOM	NO	0.645	0.245	0.020	0.379	0.061	0.281
	YES	0.645	0.245	0.020	0.379	0.061	0.281
Klarna GCN-MEAN-T	NO	0.778	0.659	0.594	0.772	0.616	0.709
	YES	0.811	0.653	0.497	0.911	0.671	0.733
Proposed Solution	NO	0.833	0.754	**0.893**	0.701	0.829	0.802
	YES	**0.869**	**0.924**	0.888	**0.920**	0.913	**0.903**